Journal of Computer and System Sciences 78 (2012) 105–118

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Markov network based ontology matching ☆

Sivan Albagli [a,b,*], Rachel Ben-Eliyahu-Zohary [c,1], Solomon E. Shimony [a]

[a] Dept. of Computer Science, Ben-Gurion University, Israel
[b] Computer Science Department, Technion – Israel Institute of Technology, Haifa 32000, Israel
[c] Department of Software Engineering, Jerusalem College of Engineering, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 11 August 2009
Received in revised form 20 August 2010
Accepted 10 February 2011
Available online 24 March 2011

Keywords:
Ontology matching
Probabilistic reasoning
Markov networks

ABSTRACT

Ontology matching is a vital step whenever there is a need to integrate and reason about overlapping domains of knowledge. Systems that automate this task are of a great need. iMatch is a probabilistic scheme for ontology matching based on Markov networks, which has several advantages over other probabilistic schemes. First, it handles the high computational complexity by doing approximate reasoning, rather then by ad-hoc pruning. Second, the probabilities that it uses are learned from matched data. Finally, iMatch naturally supports interactive semi-automatic matches. Experiments using the standard benchmark tests that compare our approach with the most promising existing systems show that iMatch is one of the top performers.

1. Introduction

Ontology matching has emerged as a crucial step when information sources are being integrated, such as when companies are being merged and their corresponding knowledge bases are to be united. As information sources grow rapidly, manual ontology matching becomes tedious and time-consuming and consequently leads to errors and frustration. Therefore, automated ontology matching systems have been developed. In this paper, we present a probabilistic scheme for interactive ontology matching based on Markov networks, called iMatch.[2]

The first approach for using inference in structured probability models to improve the quality of existing ontology mappings was introduced as OMEN in [1]. OMEN uses a Bayesian network to represent the influences between potential concept mappings across ontologies. The method presented here also performs inference over networks, albeit with several non-trivial modifications.

First, iMatch uses Markov networks rather than Bayesian networks. This representation is more natural since there is no inherent causality in ontology matching. Second, iMatch uses approximate reasoning to confront the formidable computation involved rather than arbitrary pruning as done by OMEN. Third, the clique potentials used in the Markov networks are learned from data, instead of being provided in advance by the system designer. Finally, iMatch performs better than OMEN, as well as numerous other methods, on standard benchmarks.

In the empirical evaluation, we use benchmarks from the Information Interpretation and Integration Conference (I3CON).[3] These standard benchmarks are provided by the initiative to forge a consensus for matching systems evaluation, called the

☆ This is a revised and extended version of a paper with the same title appearing in proceedings IJCAI-2009.

* Corresponding author at: Computer Science Department, Technion – Israel Institute of Technology, Haifa 32000, Israel.
E-mail addresses: albagli@cs.bgu.ac.il, al.sivan@gmail.com (S. Albagli), rbz@jce.ac.il (R. Ben-Eliyahu-Zohary), shimony@cs.bgu.ac.il (S.E. Shimony).

1 Part of this work was done while the author was at the Communication Systems Engineering Department at Ben-Gurion University, Israel.
2 iMatch stands for **i**nteractive **Ma**rcov networks matcher.
3 http://www.atl.external.lmco.com/projects/ontology/i3con.html.

0022-0000/$ – see front matter
doi:10.1016/j.jcss.2011.02.014

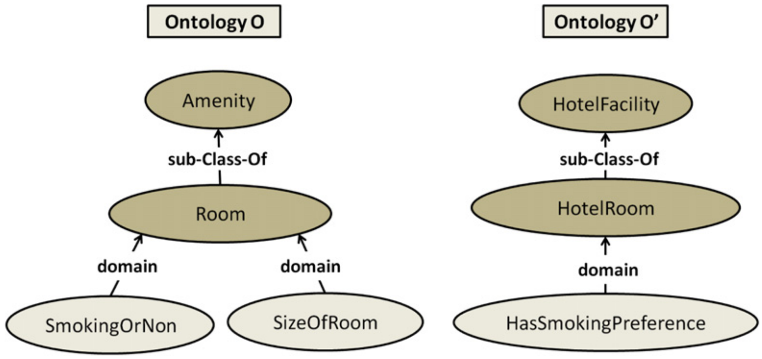

Fig. 1. Two ontologies, O and O', to be matched.

Ontology Alignment Evaluation Initiative (OAEI). The quality of the ontology matching systems can thus be assessed and compared, and the challenge of building a faithful and fast system is now much more definitive. Using the test cases suggested by OAEI, we have conducted a line of experiments in which we compare iMatch with state of the art ontology matching systems such as Lily [2], ASMOV [3], and RiMOM [4]. Our experiments show that iMatch is competitive with, and in some cases better than, the best ontology matching systems known so far. We have also experimented on an algorithm based on pseudo interaction that can be seen as an attempt to find a globally consistent alignment, and on learning more refined distribution tables. Both of these extensions are promising in that they further improve the quality of iMatch matching in most cases.

The sequel is organized as follows. We begin (Section 2) with some essential background on ontology matching and Markov networks. Then the elements of iMatch are described (Section 3), starting with the construction of the Markov network topology, the clique potentials, and reasoning over the network. This is followed by an extensive empirical evaluation of iMatch under different settings, and a comparison of matching results with some state of the art matchers (Section 4). Finally, we discuss some closely related work on matching (Section 5).

2. Background

We begin with preliminaries on ontology matching and on structured probability models.

2.1. Ontology matching

We view ontology matching as the problem of matching labeled graphs. That is, given ontologies O and O' (see Fig. 1) plus some additional information I, find the best mapping μ between nodes in O and nodes in O', where "best" is defined according to some predefined measure.

The labels in the ontology convey information, as they are usually sequences of (parts of) words in a natural language. The additional information I is frequently some knowledge base containing words and their meanings, how words are typically abbreviated, and so on. In addition, in an interactive setting, user inputs (such as partial matching indications: a human user may be sure that a node labeled "Room" in O should match the node labeled "HotelRoom" in O') can also be seen as part of I. In this paper, we focus on performing what is sometimes called a "second-line matching", that is, we assume that a probability of match (or at least some form of similarity measure) over pairs of nodes (o, o') is provided by what is called by [5] "a first-line matcher". Then, we exploit the structure and type information of the ontologies in conjunction with the information provided by the first-line matcher to compute the matching. In general, an ontology mapping μ can be a general mapping between subsets of nodes in O and subsets of nodes in O'. However, in most scenarios μ is constrained in various ways.

One commonly used **constraint** is requiring μ to be **one-to-one**. That is, for all $o \in O$, $\mu(o) \in O' \cup \{\perp\}$ (where $\perp$ stands for "o has no counterpart in O'"), and for each $o' \in O'$ there is at most one $o \in O$ such that $\mu(o) = o'$. Likewise for the partial inverse mapping μ^{-1}. This paper for the most part assumes the **one-to-one constraint**, but we also show how the proposed model can be generalized to mappings that are one to many or many to one.

For simplicity, we use a rather simple standard scheme of ontology representation, where the nodes of an ontology graph (see Fig. 1) denote entities such as classes and properties. The graph is constructed from a representation of the ontology provided for the benchmarks in a standard ontology representation language, as can be seen by the following ontology fragment in RDF, from which the left-hand side graph in Fig. 1 was constructed.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns="http://www.cs.bgu.ac.il/RDF">
```